Table 21 Summary of ML-based Misuse Detection

Ref.	ML Technique	Dataset	Features	Evaluation	
				Settings	Results
Cannady [84]	Supervised NN (*offline*)	Normal: RealSecure Attack: [143, 368]	TCP, IP, and ICMP header fields and payload	-1 Layer MLP: 9, [a], 2 -Sigmoid function -Number of nodes in hidden layers determined by trial & error	DR: 89%-91% Training + Testing runtime: 26.13 hrs
Pfahringer [358]	Supervised Ensemble of C5 DTs (*offline*)	KDD Cup [257]	all 41 features	-Two-processor (2x300Mhz) -512M memory, 9 GB disc Solaris OS 5.6 -10-folds cross-validation	DR Normal: 99.5% DR Probe: 83.3% DR DoS: 97.1% DR U2R: 13.2% DR R2L: 8.4% Training: 24 h
Pan et al. [344]	Supervised NN and C4.5 DT (*offline*)	KDD Cup [257]	all 41 features	-29,313 training data records -111,858 testing data records -1 Layer MLP: 70-14-6 -NN trained until MSE = 0.001 or # Epochs = 1500 -Selected attacks for U2L and R2L -After-the-event analysis	DR Normal : 99.5% DR DoS: 97.3% DR Probe (Satan): 95.3% DR Probe (Portsweep): 94.9% DR U2R: 72.7% DR R2L: 100% ADR: 93.28% FP: 0.2%
Moradi et al. [322]	Supervised NN (*offline*)	KDD Cup [257]	35 features	-12,159 training data records -900 validation data records -6,996 testing data records -Attacks: SYN Flood and Satan -2 Layers MLP: 35 35 35 3 -1 Layer MLP: 35 45 35 -ESVM Method	2 Layers MLP DR: 80% 2 Layers MLP Training time > 25 hrs 2 Layers MLP w/ ESVM DR: 90% 2 Layers MLP w/ ESVM Training time < 5 hrs 1 Layers MLP w/ ESVM DR: 87%
Chebrolu et al. [90]	Supervised BN and CART (*offline*)	KDD Cup [257]	Feature Selection using Markov Blanket and Gini rule	-5,092 training data records -6,890 testing data records - AMD Athlon 1.67 GHz processor with 992 MB of RAM	DR Normal: 100% DR Probe: 100% DR DoS: 100% DR U2R: 84% DR R2L: 99.47% Training BN time: 11.03 ∼ 25.19 sec Testing BN time: 5.01 ∼ 12.13 sec Training CART time : 0.59 ∼ 1.15 sec Testing CART time: 0.02∼ 0.13 sec
Amor et al. [20]	Supervised NB (*offline*)	KDD Cup [257]	all 41 features	-494,019 training data records -311,029 testing data records -Pentium III 700 Mhz processor	DR Normal: 97.68% PCC DoS: 96.65% PCC R2L: 8.66% PCC U2R: 11.84% PCC Probing: 88.33%
Stein et al. [421]	Supervised C4.5 DT (*offline*)	KDD Cup [257]	GA-based feature selection	-489,843 training data records -311,029 testing data records -10-fold cross validation -GA ran for 100 generations	Error rate DoS: 2.22% Error rate Probe: 1.67% Error rate R2L: 19.9% Error rate U2R: 0.1%
Paddabachigari et al. [354]	Supervised Ensemble of SVM, DT, and SVM-DT *Offline*	KDD Cup [257]	all 41 features	5,092 training data records 6,890 testing data records AMD Athlon, 1.67 GHz processor with 992 MB of RAM -Polynomial kernel	DR Normal: 99.7% DR Probe:100% DR DoS: 99.92% DR U2R: 68% DR R2L: 97.16% Training time: 1∼ 19 sec Testing time: 0.03∼ 2.11 sec
Sangkatsanee et al. [402]	Supervised C4.5 DT (*online*)	Normal: Reliability Lab Data 2009 (RLD09) Attack: [341, 444, 475]	TCP, UPD, and ICMP header fields	-55,000 training data records -102,959 testing data records -12 features -2.83 GHz Intel Pentium Core2 Quad 9550 processor with 4 GB RAM and 100 Mbps LAN -Platform used: Weka V.3.6.0	DR Normal: 99.43% DR DoS: 99.17% DR Probe: 98.73% Detection speed: 2∼ 3 sec

Table 21 Summary of ML-based Misuse Detection *(Continued)*

Ref.	ML Technique	Dataset	Features	Evaluation	
				Settings	Results
Miller et al. [314]	Supervised Ensemble MPML (*Offline*)	NSL-KDD [438]	all 41 features	-125,973 training records -22,544 testing records -3 NBs trained w/ 12, 9, 9 features -Platform used Weka [288]	TP: 84.137% FP: 15.863%
Li et al. [272]	Supervised TCM K-NN (*Offline*)	KDD Cup [257]	all 41 features 8 features selected using Chi-square	-Intel Pentium 4, 1.73 GHz, 1 GB RAM, Windows XP Professional - Platform Weka [288] -49,402 training records -12,350 testing records -$K = 50$	41 features: TP 99.7% 41 features: FP 0% 8 features: TP 99.6% 8 features: FP 0.1%

[a]Determined empirically, Mean Square Error (MSE), Percentage Correct Classification (PCC), Average Detection Rate (ADR), Early Stop Validation Method (ESVM)

host. Results show that the NN is able to correctly identify normal and attack records 89-91% of the time.

In 1999, the KDD cup was launched in conjunction with the KDD'99 conference. The objective of the contest was the design of a classifier that is capable of distinguishing between normal and attack connections in a network. A dataset was publicly provided for this contest [257], and since then became the primary dataset used in ML-based intrusion detection literature. It consists of 5 categories of attacks, including DoS, probing, user-to-root (U2R) and root-to-local (R2L), in addition to normal connections. The top three contestants employed DT-based solutions [421]. The winner of the contest [358] used an ensemble of 50 times 10 C5 DTs with a mixture of bagging and boosting [377]. The results of the proposed method are presented in Table 21. Clearly, the proposed approach performs poorly for U2R and R2L attack categories. The authors do mention that many of the decisions were pragmatic and encouraged more scientific endeavors. Subsequently, an extensive body of literature emerged for ML-based intrusion detection using the KDD'99 dataset, in efforts to improve on these results, where some use the winners' results as a benchmark.

For instance, Moradi et al. [322] investigate the application of NN for multi-class classification using the KDD'99 dataset. Specifically, the authors focused on DoS and probing attacks. As opposed to the work of [84], two NNs were trained: one with a single hidden layer and the second with two hidden layers, to increase the precision of attack classification. They leverage the *Early Stopping Validation Method* [366] to reduce training and validation time of the NN to less than 5 hours. As expected, the NN with 2 hidden layers achieves a higher accuracy of 91%, compared to the 87% accuracy of the NN with a single hidden layer.

Amor et al. [20] compare NB and DT also using KDD'99 dataset, and promote NB's linear training and classification times as a competitive alternative to DT. NB is found to be 7 times faster in learning and classification than DT. For whole attacks, DT shows a slightly higher accuracy over NB. However, NB achieves better accuracy for DoS, R2L, and probing attacks. Both NB and DT perform poorly for R2L and U2R attacks. In fact, Sabhnani and Serpen [398] expose that no classifiers can be trained successfully on the KDD dataset to perform misuse detection for U2R or R2L attack categories. This is due to the deficiencies and limitations of the KDD dataset rather than the inadequacies of the proposed algorithms.

The authors found via multiple analysis techniques that the training and testing datasets represent dissimilar hypothesis for the U2R and R2L attack categories; so if one would employ any algorithm that attempts to learn the signature of these attacks using the training dataset is bound to perform poorly on the testing dataset. Yet, the work in [344] reports surprisingly impressive detection accuracy for U2R and R2L. Here, a hybrid of BP NN with C4.5 is proposed, where BP NN is used to detect DoS and probing attacks, and C4.5 for U2R and R2L. For U2R and R2L only a subcategory of attacks is considered (yielding a total of 11 U2R connections out of more than 200 in the original dataset and $\sim$2000 out of more than 15000 for R2L connections). *After-the-event* analysis is also performed to feed C4.5 with new rules in the event of misclassification.

Other seminal works consider hybrid and ensemble methods for misuse detection [90, 354, 421]. The goal of ensemble methods is to integrate different ML techniques to leverage their benefits and overcome their individual limitations. When applied to misuse detection, and more specifically for the KDD'99 dataset, these work focused on looking at which ML technique works best for a class of connections. For instance, Peddabachigari et al. [354] propose an IDS that leverages an ensemble of DT, SVM with polynomial kernel based function, and hybrid DT-SVM to detect various different cases of misuse. Through